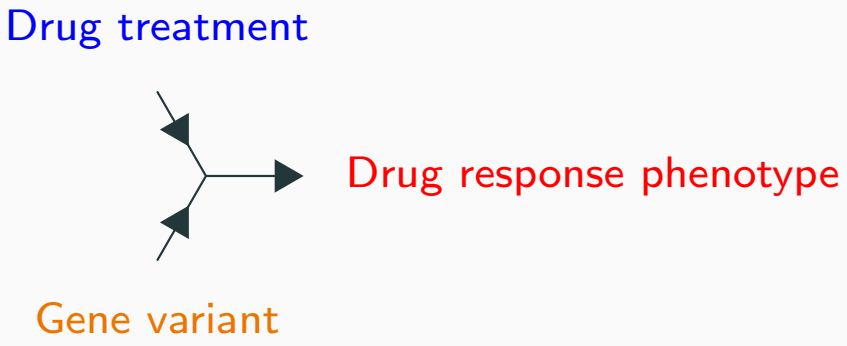
The PractiKPharma project (2016-2020)

Knowledge extraction and comparison for pharmacogenomics

October 8, 2020

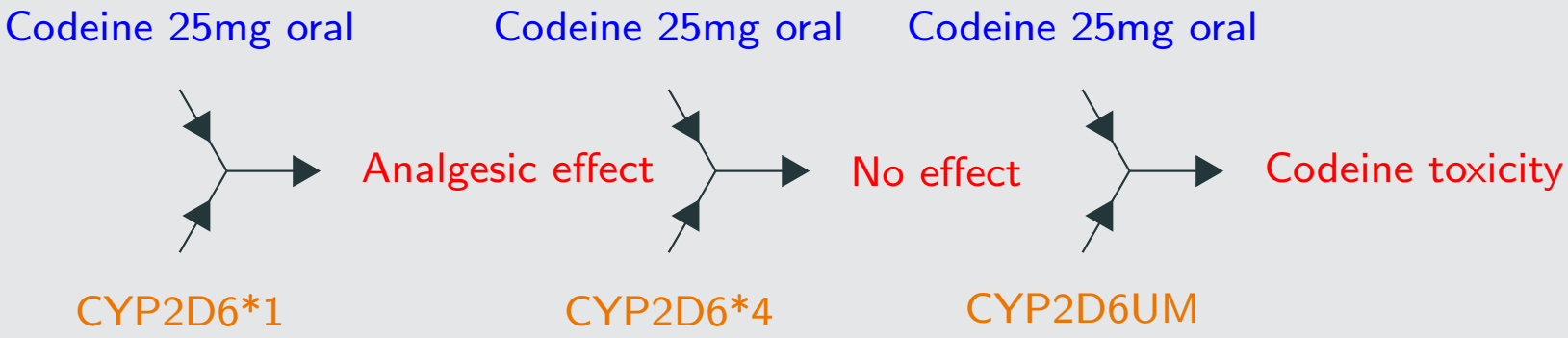
Application domain: pharmacogenomics (PGx)

PGx studies the influence of genomics on individual drug response



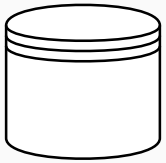
Example

CYP2D6 gene variants influence codein response

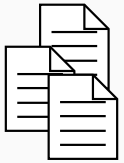


Overview of PractiKPharma

Sources of PGx relations



Specialized databases (PharmGKB)

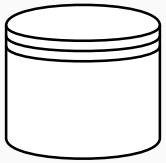


The biomedical litterature

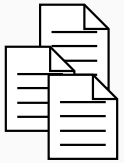
Electronic Health Records (EHRs)

Overview of PractiKPharma

Sources of PGx relations



Specialized databases (PharmGKB)



The biomedical litterature

Electronic Health Records (EHRs)

Results: 2 open source data resources

- A manually annotated corpus: **PGxCorpus**
- A PGx knowledge graph: **PGxLOD**

Knowledge extraction

Knowledge extraction from the literature

Motivation:

- Most of PGx knowledge is available in the literature
- State-of-the-art methods for relation extraction are supervised
- Few small annotated corpora exist involving PGx entities

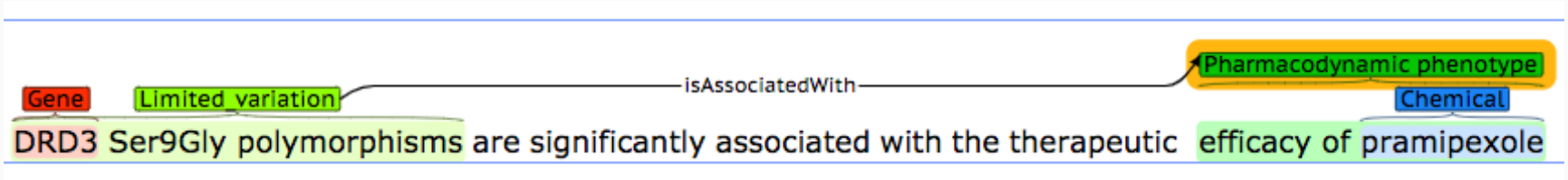
Approaches:

- Build manually a corpus for PGx relations
- Combine transfer learning and our domain specific corpus

PGxCorpus [Legrand et al., 2020]

A manually annotated corpus of **PubMed abstracts**

- all three PGx key entities annotated (Genomic factors, Drugs and Drug response phenotype)
- their relationships



Summary:

- 11 annotators involved
- 8 months
- 945 sentences, 6,761 entities, 2,875 relations

PGxCorpus: demonstrated utility

- Baseline experiments
  - Named entity recognition (NER) with a classical CNN [Collobert *et al.* 2011]
    F-measure (macro) = 71.93%
  - Relation Extraction (RE) with a Multi-Channel CNN [Quan *et al.*, 2016]
    F-measure (macro) = 54.04%

- 2019: BERT came!
  - RE with BERT + fine-tuning with PGxCorpus
    F-measure = **78.44 %**

PGxCorpus is available and open (CC A-NC 4.0)

- The corpus, annotation guidelines, code of baseline experiments: `https://github.com/practikpharma/PGxCorpus/`
- The brat server to browse the corpus: `https://pgxcorpus.loria.fr/`
- The data descriptor manuscript ([Legrand et al., 2020]): `https://www.nature.com/articles/s41597-019-0342-9`

- Useful to study other NLP task: n-ary relation extraction, discontinuous entity recognition

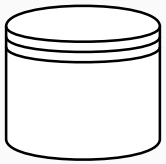
Knowledge comparison

PGxLOD: a platform for the comparison of PGx knowledge

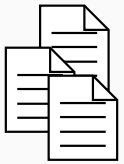
Many **PGx key entities** (genetic factors drugs, drug responses phenotypes) from public databases DrugBank, SIDER ...

Plus **PGx relations** from:

Sources of PGx relations



Specialized databases (PharmGKB)

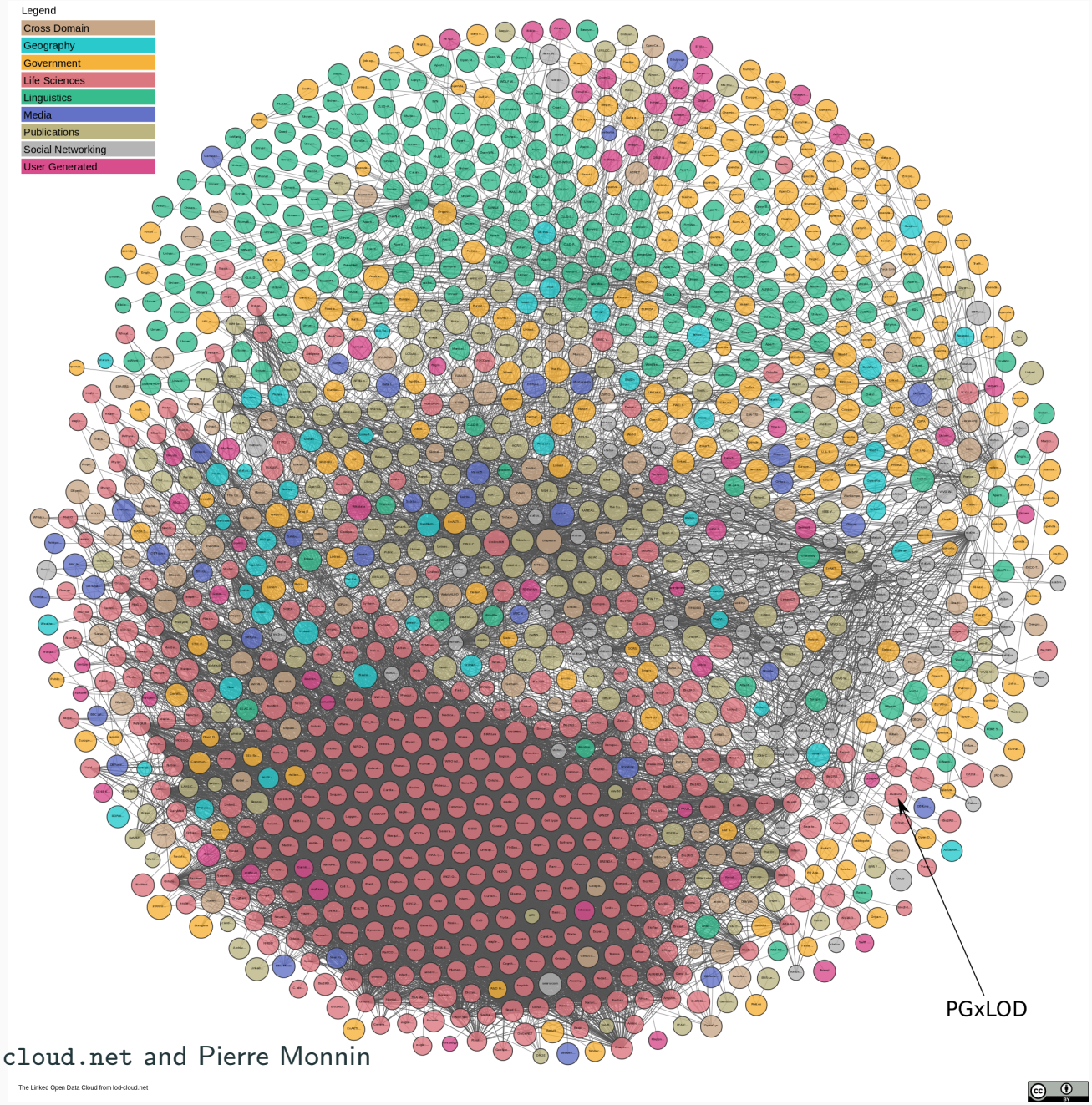


The biomedical litterature (PubMed abstracts)

EHRs mining (not automated)

Available online at `http://pgxlod.loria.fr`

PGxLOD: a platform for the comparison of PGx knowledge

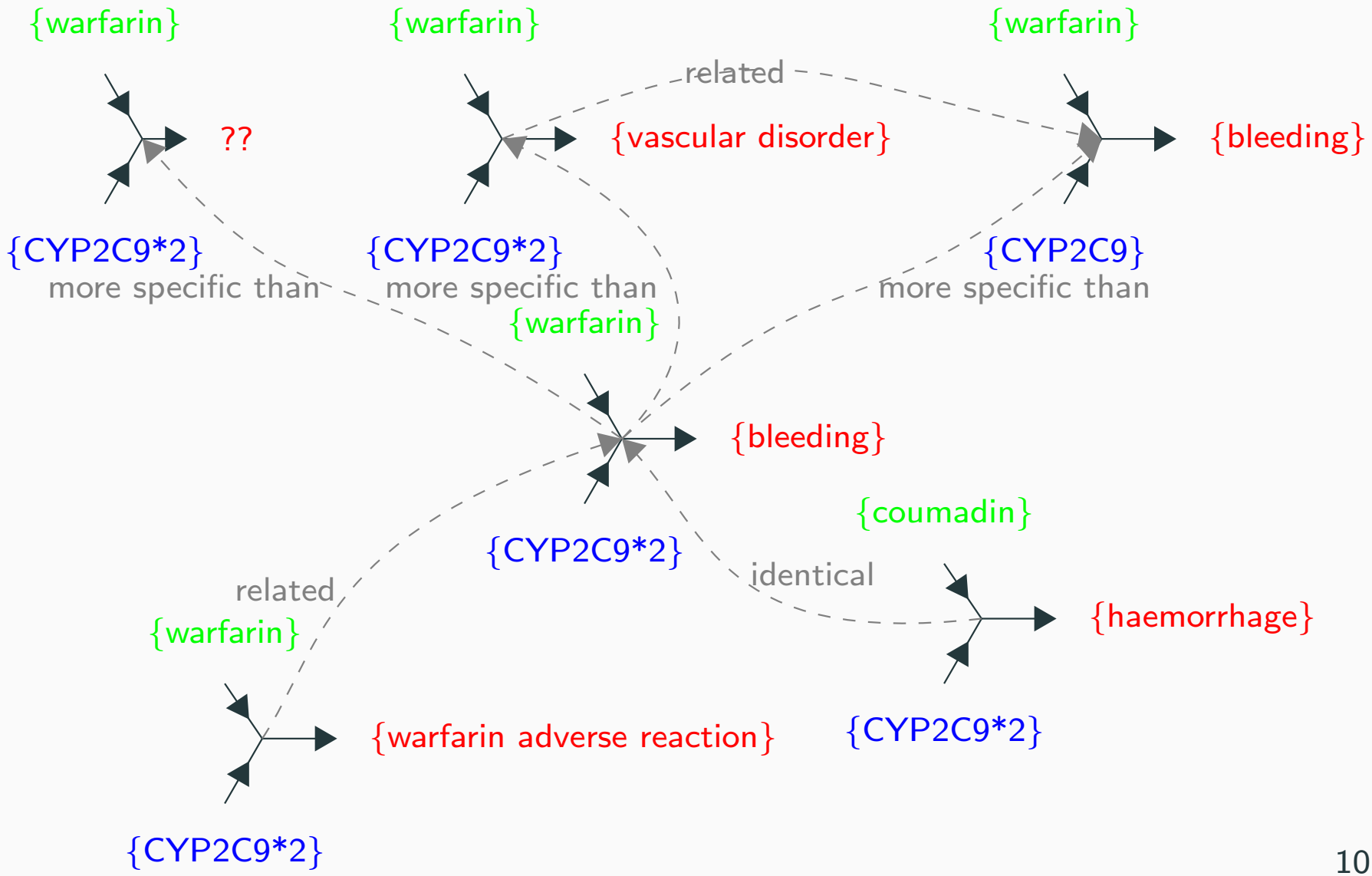


source: `lod-cloud.net` and Pierre Monnin

PGxLOD [Monnin et al., 2019a]

Concept	Number of instances
`Drug`	63,485
`GeneticFactor`	494,982
`Phenotype`	65,133
`PharmacogenomicRelationship`	50,435
from PharmGKB	3,650
from the literature	36,535
from EHR studies	10

PGx relations are heterogenously represented



Two approaches for reconciliation

1: Reconciliation rules [Monnin et al., 2020]

Example of reconciliation rule

$\forall i \in \{1, \ldots, n\}, \ \pi_i(r_1) = \pi_i(r_2) \Rightarrow \texttt{owl:sameAs}(r_1, r_2)$

2: Graph embedding for comparison of n-ary relations [Monnin et al., 2019b]

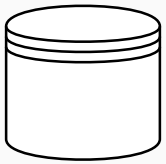
Graph Embedding [Cai et al., 2018]

converts graph structures into a d-dimensional space, in which graph properties are preserved as much a possible;

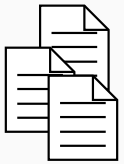
captures a similarity with more flexibility than rules, deal with noise, missing mappings ...

Perspectives on PractiKPharma

Sources of PGx relations



Specialized databases



The biomedical litterature

Electronic Health Records (EHRs)

- Still very challenging to extract knowledge automatically from EHRs

- Two promising open resources
  - **PGxCorpus**, a manually annotated corpus: `https://pgxcorpus.loria.fr/`
  - **PGxLOD**, an open knowledge graph for PGx: `https://pgxlod.loria.fr/`

Thank you for your attention

Acknowledgments

Loria, Nancy

- Joël Legrand
- Pierre Monnin
- Walid Hafiane
- Amedeo Napoli
- Chedy Raïssi
- Miguel Couceiro
- Yannick Toussaint

Lirmm, Montpellier

- Clément Jonquet
- Andon Tchechmedjiev

Hospital Georges Pompidou

- Bastien Rance
- William Digan

References i

Cai, H., Zheng, V. W., and Chang, K. C. (2018).
A comprehensive survey of graph embedding: Problems, techniques, and applications.
IEEE Trans. Knowl. Data Eng., 30(9):1616–1637.

Kipf, T. N. and Welling, M. (2016).
Semi-supervised classification with graph convolutional networks.
CoRR, abs/1609.02907.

References ii

Legrand, J., Gogdemir, R., Bousquet, C., Dalleau, K., Devignes, M.-D., Digan, W., Lee, C.-J., Ndiaye, N.-C., Petitpain, N., Ringot, P., Smaïl-Tabbone, M., Toussaint, Y., and Coulet, A. (2020).
PGxCorpus, a manually annotated corpus for pharmacogenomics.
Scientific Data , 7(3).

Monnin, P., Couceiro, M., Napoli, A., and Coulet, A. (2020).
Knowledge-Based Matching of n-ary Tuples.
In Alam, M., Braun, T., and Yun, B., editors, *25th International Conference on Conceptual Structures, ICCS 2020*, volume Lecture Notes in Computer Science of

References iii

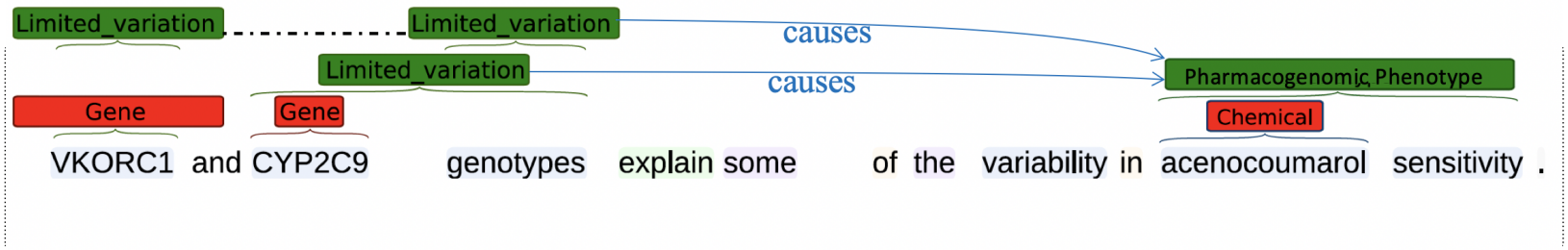
Ontologies and Concepts in Mind and Machine - 25th International Conference on Conceptual Structures, ICCS 2020, Bolzano, Italy, September 18–20, 2020, Proceedings, pages 48–56, Bolzano, Italy. Springer.

Monnin, P., Legrand, J., Husson, G., Ringot, P., Tchechmedjiev, A., Jonquet, C., Napoli, A., and Coulet, A. (2019a).
PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison.
BMC Bioinformatics, 20-S(4):139:1–139:16.

References iv

Monnin, P., Raïssi, C., Napoli, A., and Coulet, A. (2019b). **Knowledge Reconciliation with Graph Convolutional Networks: Preliminary Results.** In Alam, M., Buscaldi, D., Cochez, M., Osborne, F., Recupero, D. R., and Sack, H., editors, *DL4KG2019 - Workshop on Deep Learning for Knowledge Graphs*, volume CEUR Workshop Proceedings, Portoroz, Slovenia.

Annotation process



(1) Automatic pre-annotation

	Origin	Initial type
(1a)	*PubTator*	Chemical Disease Gene Mutation
(1b)	*PHARE*	Drug DrugMetabolite Gene GenomicRegion GenomicVariation GeneProduct Mutation Phenotype

(2) Manual annotation

Annotations 4 x 4!

First round: each sentence is annotated independentely by 2 reviewers

Second round: Annotations of each sentence are revised and merged by a third "senior" annotator

Homogeneization: Two annotators reviewed together all sentences to ensure the homogeneisation of the annotations

Application to pharmacogenomic knowledge: results

		PGKB (sd)	PGKB (ca)	Literature	EHRs
Links from	PGKB (sd)	166	0	0	0
Rule 1	PGKB (ca)	0	10,134	0	0
Encoded by	Literature	0	0	122,646	0
`owl:sameAs`	EHRs	0	0	0	0
Links from	PGKB (sd)	0	5	0	0
Rule 2	PGKB (ca)	5	1,366	0	0
Encoded by	Literature	0	0	16,692	0
`skos:closeMatch`	EHRs	0	0	0	0
Links from	PGKB (sd)	87	3	15	0
Rule 3	PGKB (ca)	9,325	605	42	0
Encoded by	Literature	0	0	75,138	0
`skos:broadMatch`	EHRs	0	0	0	0
Links from	PGKB (sd)	20	0	0	0
Rule 4	PGKB (ca)	0	110	0	0
Encoded by	Literature	0	0	18,050	0
`skos:relatedMatch`	EHRs	0	0	0	0
Links from	PGKB (sd)	100,596	287,670	414	2
Rule 5	PGKB (ca)	287,670	706,270	1,103	19
Encoded by	Literature	414	1,103	1,082,074	15
`skos:related`	EHRs	2	19	15	0

Experiments with PGxLOD

Preliminary experiments with Graph Convolutional Networks [Kipf and Welling, 2016]

Predicates ($\|\mathcal{R}\|$)	378
PGx relationships	68,686
↳ Nodes in their 3-hop neighborhood	2,943,613
↳ Edges in their 3-hop neighborhood	32,773,429
Similarity links between PGx relationships	283,248
↳ `owl:sameAs` links	109,226
↳ `skos:broadMatch` links	136,264
↳ `skos:relatedMatch` links	37,758

- Training set: $\frac{2}{3}$ – Test set: $\frac{1}{3}$
- 3-layer network
- Embeddings in $\mathbb{R}^{10}$
- Training during 60 epochs with Adam optimizer